Chained System: A Linear Combination of Different Types of Statistical Machine Translation Systems

Takako Aikawa
Microsoft Research
One Microsoft Way
Redmond, WA, 98052, USA
takakoa@microsoft.com

Achim Ruopp
Butler Hill Group
P.O. Box 935
Ridgefield, CT 06877, USA
achim@digitalsilkroad.net

Abstract

The paper explores a way to learn post-editing fixes of raw MT outputs automatically by combining two different types of statistical machine translation (SMT) systems in a linear fashion. Our proposed system (which we call a chained system) consists of two SMT systems: (i) a syntax-based SMT system and (ii) a phrase-based SMT system (Koehn, 2004). We first translate source sentences of the bi-text training data into a target language, using the syntax-based SMT. This provides us the monolingual parallel data that consist of the raw MT outputs and their corresponding human translations. We then build a phrase-based SMT system, using the monolingual parallel corpus. Our system is thus a chain of a syntax-based SMT system and a phrase-based SMT system. The benefit of the chained system is to learn post-editing fixes automatically via a phrase-based SMT system (Simard, et al., 2007a/b). We investigated the impact from the chained system on the initial SMT system in terms of BLEU, using typologically different language pairs. The results of our experiments strongly indicate that the second part of the chained system can compensate the weaknesses of the initial SMT system in a robust way by providing human-like fixes.

1 Introduction

The quality of an SMT system has improved quite a lot since late 90's and different types of SMT systems have been proposed over the last decade. The quality of SMT, however, is still not sufficient for actual use. For instance, we have been using a syntax-based SMT system for the last several years to localize technical manuals or documents. However, most of our clients end up handing off raw MT output to human post-editors due to their concerns about the quality of the MT output. Human post-editing of raw MT output is as costly as human translation from scratch. This in fact devalues the use of MT for localization.

When comparing raw MT outputs and their human post-edited translations, we often find repetitious changes of wrongly translated phrases to correct ones. This motivates a so-called "automatic post-editing" (APE) proposed by Simard et. al., (2007 a/b). In their view, the task of post-editing is considered as the task of finding mappings between raw MT output and human post-edited translation. They used a phrase-based SMT system to learn such mappings and apply it to the output of a rule-based MT system. They show impressive results by adding this phrasal system to their rule-based MT system. We took their insight and applied it to our syntax-based SMT.

Our system consists of two SMT systems: (i) a syntax-based SMT (called "treelet") system (Quirk, et al., 2005) and (ii) a phrase-based SMT system modelled on Pharaoh (Koehn, 2004). We call it "the chained system" throughout the paper. We compare the baseline treelet SMT and the chained system in terms of BLEU (Papineni, et al., 2002), using typologically different language pairs (English->Spanish, German, and Japanese). The results from our experiments show that the idea of

APE discussed in Simard et al., 2007(a/b) works for an SMT system and that it can provide human-like post-editing fixes across different language pairs automatically.

The organization of the paper is as follows. Section 2 provides an overall architecture of our chained system. Section 3 provides the design of our experiments and their results. In Section 4, we provide the linguistic error analyses of the results from our experiments. Section 5 provides our concluding remarks and future work.

2 Architecture

2.1 Training Time Overview

The overall architecture of the training of the chained system (using the English -> Spanish language pair as an example) is provided in Figure 1.

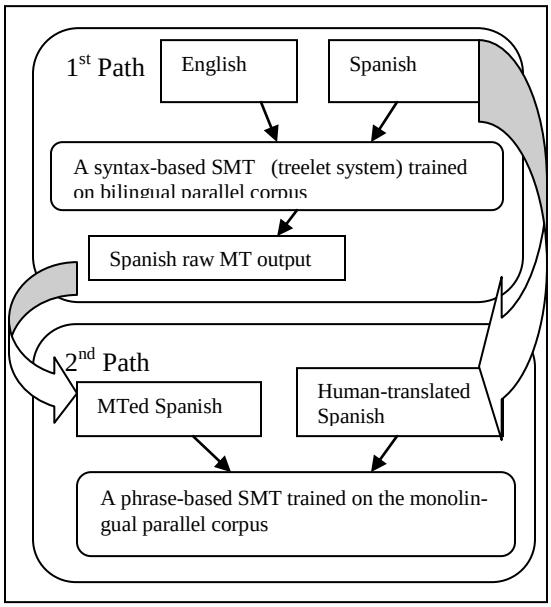

Figure 1: Training Time Overview

In the following subsections, we walk through the training process step-by-step while providing brief descriptions of the two SMT systems.

2.2 First Path: Creating Monolingual Parallel Corpus

Our baseline system (the treelet system) is a syntactically informed SMT and it requires bilingual parallel corpus data as its training data (see Quirk, et al., 2005 for technical details). We use this treelet system as the first path of the chained system.

To illustrate the training process of the chained system step-by-step, let us assume that we are going to create an English-to-Spanish chained system. As the first path of the chained system, we train the treelet system on the bilingual (English-Spanish (ES)) parallel corpus and then translate all the source English sentences of the training data into Spanish (see 1st Path in Figure 1). This gives us a new set of parallel data consisting of: (i) the Spanish MT outputs and (ii) their corresponding human translations (from the ES training data). The first phase of the chain system can be considered to be a process to create the monolingual parallel corpus via an existing SMT system, which will be used to train the second phase of the chained system.

2.3 Second Path: Training a Phrase-based System

As the second path of the chain, we train a phrase-based SMT system using the monolingual parallel corpus mentioned above. This second path is expected to learn the post-editing fixes for the raw MT output of the initial SMT. The phrase-based system we used in this paper is a re-implementation of the Pharaoh system (Koehn, 2004). The word alignment of our phrase-based system is done by an HMM-based word alignment algorithm (He, 2007). As in Koehn (2004), word alignment is performed bi-directionally; (i) from the source (the raw SMT output) to the target (the human translation of the target side of the training data) and (ii) from the target to the source. These two alignments are combined to form the final word alignment with the heuristics described in Och and Ney (2000). From this, we extract phrasal translation pairs that are consistent with the word alignment. For our experiments, we set the maximum phrase length to 4, and the maximum reordering limit for decoding to 2.

2.4 Run-time Overview

The overview of the run-time process of the chained system is provided in Figure 2.

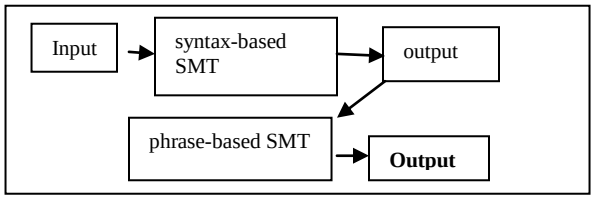

Figure 2: Run-Time Overview

The run-time process is simply the concatenation of the baseline syntax-based SMT trained on the bilingual parallel corpus (see Section 2.2) and the phrase-based system trained on the monolingual parallel data (see Section 2.3). That is, first translate an input English sentence into Spanish using the treelet system and then "re-translate" the output into Spanish using the monolingual parallel data trained phrase-based SMT system.

3 Experiments

We trained chained systems for three typologically different language pairs; (i) English->Spanish (ES), (ii) English->German (EG) and (iii) English->Japanese (EJ). For the ES system, we used data from the publicly available Europarl corpus v2 (Koehn, 2005).[1] For the EG and the EJ systems, we used technical domain data sets. Table 1 provides the list of the training, dev and test data sets.

Dataset	#sent pairs	word tokens	
ES		English	Spanish
Training	800K	16,760K	17,163K
Dev	2K	42K	43K
Test	5K	106K	109K
EG		English	German
Training	2M	29,439K	30,398K
Dev	2K	32K	31K
Test	1892	32K	34K
EJ		English	Japanese
Training	2M	29,509K	35,540K
Dev	2K	33K	39K
Test	4355	76K	95K

Table 1: Data sets (K=rounded to the nearest thousand, M=rounded to the nearest million)

[1] We limited sentences to those with the length of 2000 characters and removed XML tags as well as empty lines with their correspondences.

For each of these data sets we trained a baseline treelet system and evaluated the baseline system on the test set using BLEU.[2]

For the evaluation of the chained system, we first translated the training and dev sets using the baseline system. With these translated sets, we trained the phrase-based system. In testing, we translated the test data using the chained system (i.e., first, the baseline treelet system and then, the phrase-based system) and measured its quality using BLEU.[3]

The results are shown in Table 2. For comparison purposes we also trained a baseline phrase-based system with the same configuration as the chained phrase-based system and the bilingual dataset of the baseline system. All BLEU scores are reported in percentage.

	Baseline Treelet System	Chained System	Baseline phrase-based SMT
ES	33.17	34.28 (+1.11)	32.63
EG	46.83	50.21 (+3.38)	46.24
EJ	39.92	43.60 (+3.68)	30.28

Table 2: BLEU scores of baseline treelet, chained phrasal and baseline phrasal (the delta indicated above is the comparison between the baseline SMT system and the chained system)

As shown, all the chained systems show BLEU gains over the baseline treelet and the baseline phrase-based systems.[4] This shows that the chained system works across different language pairs. It is not yet clear to us, however, why we gained only 1 point for the ES system whereas for the EG and the EJ systems, we gained more than 3 points. One speculation for this is that the chained system works better when the domain is specified or narrower. This is a question that we plan to address in future work.

[2] We used the standard NIST BLEU scoring tool mteval-v11b.pl to obtain BLEU scores.

[3] Our test data has only one reference per sentence.

[4] We speculate that the low score for the baseline EJ phrase-based system (compared to the baseline treelt system) is because these two languages have totally opposite word orders: Japanese is head-final whereas English is head-initial. We set the distortion length to 2 for our phrase-based system. This presumably makes it hard for the phrase-based system to handle differences in order between these two languages.

4 Error Analyses

4.1 Method

To investigate the impact of the chained system from linguistic perspectives, we conducted error analyses on the results of the chained systems mentioned in Section 3. To this end, we first calculated sentence-level BLEU and character edit rate (Levenshtein, 1965) over the two sets of results: (a) the outputs from the treelet system and (b) those from the chained system. Second, we calculated the differences in the two metrics between (a) and (b).[5] We assumed that the differences in the scores reflect the magnitude of the positive/negative impact from the chained system on the baseline system. We extracted top 100 positive examples (i.e., top 100 examples that have a positive value for the chained system) and bottom 100 negative ones (i.e., bottom 100 examples that have a negative value against the chained system) for each of the three language pairs. We asked the speakers of these languages to analyze types of linguistic fixes/errors that the chained system has made on these 200 examples. In the following subsections, we describe some details of our error analyses.

4.2 Positive Impact

Table 3 provides categorical descriptions of the positive changes from the chained system for all the three language pairs.

	Categorical Descriptions
ES	inflections; agreements; pronouns; negations; better lexical choices; etc.
EJ	inflections; case-markers; etc.
EG	compound nouns; better fluency; inflections; etc.

Table 3: Categorical Descriptions of the Positive Changes

4.2.1 English->Spanish

Examples (1) and (2) illustrate some of the fixes that the chained system provided for the baseline ES system.

(1) [Source] This is why we are criticizing the pressure that they are under.

a. [The baseline system output]
Por eso **nos** critican la presión **que están bajo**.
'This is why (they) are criticizing us the pressure that are low.'

b. [The chained system output]
Por este motivo criticamos las presiones **que sufren**.
'This is why (we) are criticizing the pressure that (they) suffer.'

(2) [Source] In fact, all of these religions are outlawed and have no legal status.

a. [The baseline system output]
De hecho, todas estas **religiones** están **proscrita** y han **ningún** estatuto jurídico.
'In fact, all are the religions (be) proscribed and have none legal status.'

b. [The chained system output]
De hecho, todas estas **religiones** están **prohibidas** y **no** gozan de estatuto jurídico.
'In fact, all are the religions (be) prohibited and do not enjoy (of) legal status.'

The contrast between (1a) and (1b) describes the better handling of Spanish pronouns and lexical choices. Spanish is a pro-drop language and the subject pronoun can be dropped freely. The presence of the overt pronoun *nos* ('us') in (1a) gives the wrong pronominal interpretation of "This is why (they) are criticizing *us*...", whereas the absence of an overt pronoun in (1b) gives us the correct pronominal interpretation of "This is why (we) are criticizing....". The contrast in the lexical choices between *que están bajo* ('that be below') in (1a) and *que sufren* ('that suffer') in (1b) indicates that the chained system provided the better lexical choice in this context as well.

The contrast between (2a) and (2b) illustrates the improvements brought by the chained system in terms of agreement and negation. In (2a), there is no agreement between the noun *religiones* 'religions' (which has +feminine and +plural) and its modifying adjective *proscrita* 'proscribed' (which has +feminine and +singular), resulting in the mismatch in number. In (2b), on the other hand, the noun *religiones* agrees with its modifying adjective *prohibidas* 'prohibited' both in gender and number. (2a) is worse than (2b) in terms of the handling of negation as well: in (2a), the so-called indefinite negative adjective *ningún* '(not) any' occurs, which requires a (true) negation marker. But (2a) does not have it, resulting in ungrammaticality. In (2b), by contrast, the chained system

[5] We calculated a combined metric between these two measures in the following way: diff scores = (sentBLEU(chained)-sentBLEU(baseline))* characterEditRate.

provided *no* 'no' and hence, the negation is nicely recovered.

4.2.2 English -> Japanese

Example (3) below illustrates the inflection fix provided by the chained system. In (3a), the part 記載 修正 プログラム'(lit) the hotfix describe' is missing the light verb する 'to do', which leads to the wrong interpretation. The chained system nicely supplied the missing predicate (i.e., the underline part in (3b)), resulting in the correct translation of 'the hotfix described in this article'.

(3) [Source] Install the hotfix described in this article.
a. [The baseline system output]
この 資料に*記載* 修正プログラムをインストールします。
'(lit.) Install the describe hotfix in this article.'
b. [The chained system output]
この資料に*記載されている*修正プログラムをインストールします。
'(lit.) Install the described hotfix in this article.'

Other prototypical fixes that the chained system provided for the baseline EJ system are those for case-markers. Japanese is a word order-free language. To specify the argument structure of a predicate (e.g., the subject, the object, etc.), it requires a case-marker for nouns. In (4a), the noun, *Windows XP*, serves as the object of the verb *start* and it requires the object case-marker を. The lack of the case-marker in (4a) makes the translation ungrammatical. In (4b), by contrast, the chained system nicely supplies the object marker を, resulting in the correct translation.

(4) [Source] If you cannot start Windows XP
a. [The baseline system output]
Windows XP 起動できない場合
'if (you) cannot start Windows XP'
b. [The chained system output]
Windows XP を起動できない場合
'if (you) cannot start Windows XP'

4.2.3 English -> German

As for the EG chained system, it provided many nice fixes for the treatment of German compound nouns as shown in (5)-(6).

(5) [Source] Color Mode
a. [The baseline system output]
Farbe Modus (not correct)
b. [The chained system output]
Farbmodus

(6) [Source] Property Tab Reference
a. [The baseline output]
Eigenschaft Registerkarte Referenz (not correct)
b. [The chained system output]
Registerkartenreferenz

Another type of fix that the EG chained system provided is to improve the fluency by changing the word order or by correcting phrasal expressions. For instance, (7b) below is much more fluent than (7a).

(7) [Source] Supplementary item overview (report)
a. [The baseline system output]
zusätzliche Artikel Übersicht (Bericht)
additional item overview report
b. [The chained system output]
Übersicht über zusätzliche Artikel (Bericht)
overview over additional items report

4.3 Negative Impact

We also examined the negative cases to investigate what types of damage the chained system has done to the output of the baseline system. One of the most common errors that we found in the chained system output across all these language pairs is that the chained system sometimes deletes content words. Table 4 provides some of such examples.

	English	Baseline	Chained
ES	There is no easy one-way ticket back.	**Existe** no fácil billete unidireccional atrás.	No fácil billete unidireccional atrás. <= the main predicate is missing.
ES	They will not do it!	No **harán que** !	no ! <=the main predicate is missing.
EJ	Arc Serve backup client	円弧 **Serve** バックアップ クライアント ('Arc Serve backup client')	円弧バックアップ クライアント ('Arc backup client') <=*Serve* is missing.
EJ	Class1.cs is created by default.	デフォルト で Class1.**cs** が 作成 されます。('Class1.cs is created by default.')	デフォルト で Class1 が 作成 されます。('Class1 is created by default.') <=.*cs* is missing.
EG	This is the standard output:	Dies ist die **Standard** -Ausgabe :	Dies ist die -output : <=Standard is missing.

Table 4: Missing Content Words (the bold-faced parts indicate the missing parts from the outputs of the chained systems)

Another prototypical error that the chained system made is adding extra information. Table 5 provides samples of such examples.

	English	Baseline	Chained
ES	That is the financial perspective on which this Parliament would like to vote.	Es las perspectivas financieras en el que este Parlamento desearía votar.	Es las perspectivas financieras en el que este Parlamento desearía votar favorablemente.
EJ	... Computer Associates eTrust 7.0 to Windows XP SP2 fails.	..Computer Associates eTrust 7.0 から、リモート インストール が 失敗します。	..Computer Associates eTrust Antivirus 7.0 から、リモート インストール が 失敗します。
EG	To use basic find to search for text	mit einfachen Suchen nach Text suchen	Grundformen Suchen nach Text suchen

Table 5: Adding Extra Information

The underlined parts of the chained system outputs in Table 5 are the extra information that does not exist in the source English sentences. Although the number of such negative cases (e.g., those in Table 4 and Table 5) is not large, these types of errors should not be introduced by the chained system.

5. Concluding Remarks/Future Work

The proposed system is a chain of two types of SMT systems. In this paper, we used the syntax-based system as our initial SMT system and trained a phrase-based SMT system based on the monolingual parallel corpus data created by the initial system. The results from our experiments strongly indicate that the chained system works for typologically different languages and it can provide a big boost over the overall quality of the initial SMT system. The paper also gives strong support for the idea of creating an APE system using a phrase-based system, which is entertained by Simard et al. (2007a/b).

There are several things that we would like to investigate in the future. First, as mentioned in Section 4, the chained system sometimes deletes content words from the raw SMT output. Also, it adds additional word(s) to the raw SMT output. We would like to investigate these cases further, so that we can prevent such errors automatically by blocking certain phrasal mappings from our phrase-based SMT system. Second, we would like to investigate if translating the training data with an SMT system that was trained on the same data leads to overfitting problems for our chained system. Third, we would like to see whether the proposed approach works for an SMT system other than a syntax-based SMT (e.g., a treelet system). For instance, we can build easily a chained system that consists of two phrase-based systems. We would like see whether such a chained system would give a similar boost as the proposed chained system did. Last, we would like to examine further what decoder settings (e.g., maximum phrase length, distortion length, etc. of a phrase-based SMT system) would work best for a chained system.

References

Xiaodong He. 2007. *Using Word Dependent Transition Models in HMM based Word Alignment for Statistical Machine Translation*. ACL Workshop on Statistical Machine Translation

Philipp Koehn. 2005. *Europarl: A Parallel Corpus for Statistical Machine Translation.* MT Summit

Philipp Koehn. 2004. *Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models.* AMTA 2004

Vladimir I. Levenshtein. 1965. *Binary codes with correction for deletions and insertions of the symbol 1*. Problemy Peredachi Informacii

Franz Josef Och and Hermann Ney. 2000. *Improved statistical alignment models.* Proceedings of the ACL.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation.* IBM Research Division, Thomas J. Watson Research Center, Technical Report RC22176 (W0109-022).

Chris Quirk, Arul Menenzes, and Colin Cherry. 2005. *Dependency Treelet Translation: Syntactically Informed Phrasal SMT.* Proceedings of ACL. Ann Arbor, MI.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007a. *Statistical Phrase-based Post-editing.* NAACL-HLT. Rochester, New York.

Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. 2007b. *Rule-based Translation With Statistical Phrase-based Post-editing.* ACL 2007.